Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments

Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments

Frank Schumann
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

Wolfgang Einhäuser-Treyer
Department of Neurophysics, Philipps-University Marburg, Marburg, Germany

Johannes Vockeroth
Chair for Clinical Neurosciences, University of Munich Hospital, Munich, Germany

Klaus Bartl
Chair for Clinical Neurosciences, University of Munich Hospital, Munich, Germany

Erich Schneider
Chair for Clinical Neurosciences, University of Munich Hospital, Munich, Germany

Peter König
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany

During free exploration, humans adjust their gaze by combining body, head, and eye movements. Laboratory experiments on the stimulus features driving gaze, however, typically focus on eye-in-head movements, use potentially biased stimuli, and restrict the field of view. Our novel wearable eye-tracking system (EyeSeeCam) overcomes these limitations. We recorded gaze- and head-centered videos of the visual input of observers freely exploring real-world environments (4 indoor, 8 outdoor), yielding ~10 h of data. Global power spectra reveal little difference between head- and gaze-centered recordings. Local stimulus features exhibit spatial biases in head-centered coordinates, which are environment-dependent, but consistent across observers. Eye-in-head movements center these biases in gaze-centered coordinates, leading to elevated "salient" features at center of gaze. This shows that central biases in image feature distributions in "natural" photographs are not a property of environments, but of stimuli already gaze-centered by the photographer. Further central biases in laboratory subjects' fixation distributions do not result from re-centering of the eyes but are an artifact of display restrictions. Hence, our findings demonstrate that the concept of feature "saliency" transfers from the laboratory to free exploration, but also highlight the importance of experiments with freely moving eyes, head, and body.

Keywords: eye movements, natural scenes, saliency, attention, natural behavior, human, eye tracking

Citation: Schumann, F., Einhäuser-Treyer, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*(14):12, 1–17, http://journalofvision.org/8/14/12/, doi:10.1167/8.14.12.

Introduction

During natural behavior, humans and other primates move their eyes to shift gaze approximately 3 to 5 times per second. In addition to this volitional adjustment of the focus of attention and spatial resolution, a wide range of compensatory movements are employed to stabilize gaze during ego-motion or to track moving objects. In the context of natural stimuli, this raises a variety of research questions that have been addressed using a wide range of techniques. In particular: to what degree do stimulus features drive volitional gaze shifts, what is their relation to attention, how do different eye-movement systems

doi: 10.1167/8.14.12 Received January 15, 2008; published October 23, 2008 ISSN 1534-7362

interact under various conditions and what are the roles of environment and task?

A large body of studies deals with the allocation of gaze as measure of so-called "overt" attention. In a typical setting, dating at least back to Buswell's (1935) seminal work, observers are shown photographs on paper or on a computer screen while their eye-position is tracked. Many studies show that fixation probability on natural images correlates with low-level features such as luminance contrast (Reinagel & Zador, 1999), edge density (Baddeley & Tatler, 2006; Mannan, Ruddock, & Wooding, 1996, 1997), and texture contrast (Einhäuser & König, 2003; Parkhurst & Niebur, 2004). Bottom-up models of attention such as Koch and Ullman's (1985) saliency map can, to some extent, predict gaze allocation exclusively by stimulus features (Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005). The *causal* role of low-level features in guiding attention has, however, been challenged (Einhäuser & König, 2003), suggesting a decisive role of higher-order correlations (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000), contextual cues (Torralba, Oliva, Castelhano, & Henderson, 2006), and objects such as faces (Cerf, Harel, Einhäuser, & Koch, 2008). Along similar lines, informative image regions are preferentially fixated (Kayser, Nielsen, & Logothetis, 2006), and interesting objects in turn correlate with low-level saliency in natural scenes (Elazary & Itti, 2008). Similarly, early studies already demonstrated the importance of task on eye movements (Buswell, 1935; Yarbus, 1967), to an extent that bottom-up models may lose all their predictive power (Henderson, Brockmole, Castelhano, & Mack, 2006; Rothkopf, Ballard, & Hayhoe, 2007) and bottom-up cues are immediately overruled or even reversed (Einhäuser, Rutishauser, & Koch, 2008). Recently, combining bottom-up saliency with top-down biases has improved performance in the modeling of search tasks (Hamker, 2006; Navalpakkam & Itti, 2007; Pomplun, 2006; Rutishauser & Koch, 2007). Although extensions to interactive scenarios have been proposed (Peters & Itti, 2008), such models are typically tested in head-restrained laboratory settings for technical reasons, which are restricted to eye-in-head movements, involve a potentially biased choice of stimuli, and present stimuli in a limited field of view. This demands models of gaze allocation to be tested in less restrained settings.

A number of pioneering studies approached eye-movements in real-world settings. Eye-movement patterns were characterized in a variety of every-day activities such as making tea, preparing food (Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999), or washing hands (Pelz & Canosa, 2001). Other studies investigated the support of eye-movement for highly skilled activities such as throwing or catching a ball (Hayhoe, Mannie, Sullivan, & Gorgosm, 2005), playing squash (Chajka et al., 2006) or cricket (Land & McLeod, 2000), piano sight-reading (Furneaux & Land, 1999), or artists sketching a portrait (Mial & Tchalenko, 2001). Recently, the effect of different realistic (navigation) tasks on eye movements was quantified in a virtual environment setup (Rothkopf et al., 2007). These studies allowed important insight into gaze allocation during natural behavior. Most importantly, eye-movements are highly influenced by task constraints, demonstrating that humans use learned internal knowledge about the actions performed to actively pick up contextually important information at current and anticipated points of action (Ballard, Hayhoe, Li, & Whitehead, 1992; Hayhoe & Ballard, 2005; Land, 2006). On the other hand, during bodily motor control, human eye-movements can be actively employed for interpretation and control of ecological motor-control variables such as optic flow patterns or the angular directions of external reference points relative to one's own body (Glasauer, Schneider, Jahn, Strupp, & Brandt, 2005; Harris & Rogers, 1999; Land & Lee, 1994; Lappe, Bremmer, & van den Berg, 1999a, 1999b; Perrone & Stone, 1994; Wilkie & Wann, 2003). Notwithstanding the important conclusions from those studies, they are typically restricted to very specific tasks or environments. As our interest here is to examine the relation of low-level stimulus features to gaze allocation under natural conditions, a complementary approach is required that minimizes task constraints and allows comparing free viewing in the laboratory with natural exploration. Since this approach comes at the cost that statistical statements can be made only over a data set, it requires large amounts of data for each real-world scenario and environment to be considered.

Two major issues bedevil the use of head-fixed laboratory setups for investigating gaze allocation under natural conditions. First, the contribution of human head and body movements is obviously neglected. Second, stimuli—especially those recorded by human observers—may exhibit spatial biases. If the setup constraints (initial fixation, limited screen width, etc.) additionally induce spatial biases on fixation, the contribution of local stimulus features might be misinterpreted unless corrected for this double bias (Einhäuser & König, 2003; Mannan et al., 1996; Tatler, Baddeley, & Gilchrist, 2005). Recently, Tatler (2007) demonstrated that in laboratory head-restrained conditions, central fixation biases prevail irrespective of biases in typical stimulus features. Is the central fixation bias a consequence of yet unknown stimulus biases, of resetting the eyes to a default position in their orbits, or of the artificial setup of watching stimuli on a computer screen? If one of the latter two, this would provide further evidence that stimulus statistics at gaze reported in lab experiments can arise due to a correlation of central feature biases with setup-induced central fixation biases rather than causally. It is also unclear whether an elevation of local features at gaze, causal or correlative, would transfer to real-world conditions without the restraints in the setup, i.e., where eye, head, and body can move freely. A head-restrained setting cannot resolve these issues.

By using a novel wearable recording setup (EyeSeeCam), we simultaneously recorded gaze-centered and head-

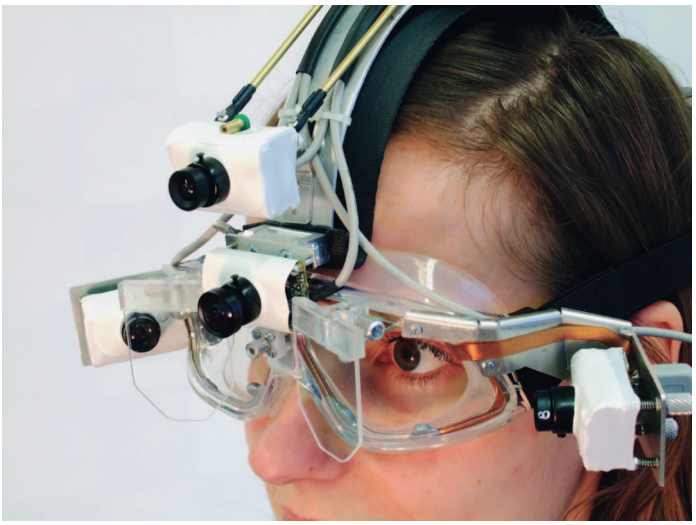


Figure 1. Setup. Movie depicting the EyeSeeCam in operation. The gaze camera follows the direction of the eye with virtually no delay. Since oculomotor compensation mechanisms are directly exploited by the EyeSeeCam, the gaze camera is stable relative to the world during steady gaze, even if the observer makes large and fast head movements.

centered videos during free exploration of various natural environments. We collected about 10 hours of data in 12 different real-world environments and analyzed the spatial distribution of stimulus features (luminance contrast, texture contrast, color contrasts, edge features, etc.) at a broad range of spatial scales. These unique data allowed us to dissociate between the contribution of eye movements (eye-in-head) and head movements (head-in-world) to stimulus statistics at the center of gaze without restraints or setup-induced biases. In particular, we quantify the extent to which eye-movements actively center salient stimulus features under natural exploration conditions, as compared to stimulus biases already present in head-centered coordinates.

Methods

Setup

Gaze- and head-centered video streams of the natural visual exploration were recorded with a custom-made wearable eye tracker (EyeSeeCam, University of Munich; Figure 1). The details of this novel system have been described previously (Brandt, Glasauer, & Schneider, 2006; Schneider et al., 2005, 2006; Vockeroth, Bardins, Bartl, Dera, & Schneider, 2007). In brief, the orientation of the eye-in-head is tracked at 192 Hz via high-speed cameras that are attached to swimming goggles worn by the observer (Figure 2). The eye position signals control the direction of a gaze-driven pivotable video camera (Firefly MV, Point Grey Research, Canada; "gaze-camera") and align it to the direction of gaze in near real time. The delay between eye movements and corresponding camera movements amounts to 26 ms. A detailed

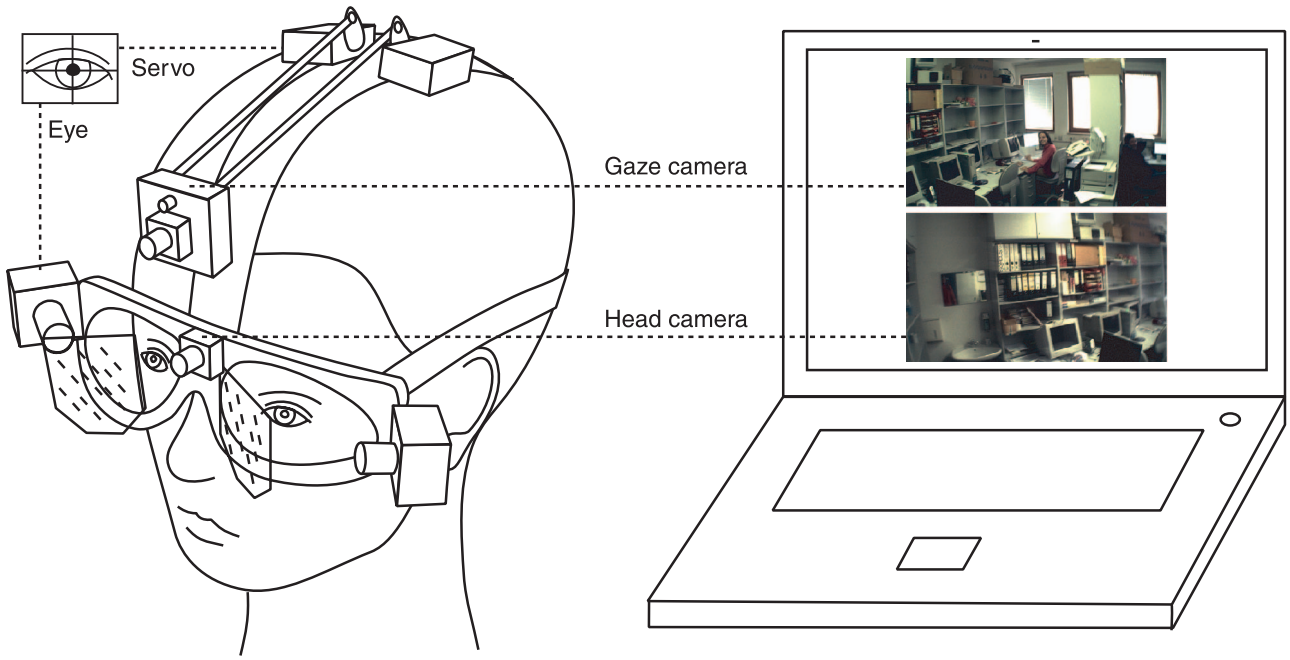


Figure 2. Schematic view. The EyeSeeCam system is mounted on swimming goggles and a lightweight mid-sagittal flat spring. A mobile laptop performs online eye tracking on the laterally attached tracker cameras and controls the servo motors to adjust the gaze-aligned camera synchronous to eye movements. The laptop records the uncompressed video signals from gaze and head camera and also serves as power supply.

evaluation of the system and its applicability to the examination of visual exploration under natural conditions is given in our previous report (Einhäuser et al., 2007).

A second camera ("head camera") is fixed with respect to the head. To allow a comparison of gaze and head images, both cameras are identical and use identical optical lenses, operate at the same resolution of 752 × 432 pixels and with the same frame rate of 25 Hz. The uncompressed and Bayer-coded video data are directly streamed to hard disk via an IEEE1394 interface. Upon request, these movies are available from the authors (Figure 3 for examples). The setup weighs 0.75 kg and therefore has only a negligible impact on human head movements.

Just like any other eye tracker, EyeSeeCam needs to be calibrated to establish a meaningful mapping between the position of the pupil in eye tracker camera coordinates and the alignment of the gaze camera. For this purpose, a small laser module is attached to the gaze camera (Figure 2). The laser projects a small but clearly visible dot onto an object like a distant wall. When the observer looks at the dot, the observer's and the camera's line of sight are almost parallel. The observer is asked to fixate and follow the laser dot, while the camera moves to 25 predefined orientations. The pupil position values are then mapped to the motor commands of the predefined camera orientations by two 3rd order polynomials. The parameters of these functions are obtained from a linear fit. During normal operations, these mapping functions are used to calculate appropriate motor commands from the pupil position.

Subjects/environments

A participant pool of seven observers (five males, two female; ages 25–40) participated in the experiment. Four (D, E, F, J) were authors. All had normal or corrected-to-normal vision and were accustomed to wearing the recording equipment. Experiments were conducted in 12 different environments (Figure 3), with two subjects per environment. Environments were selected to achieve a variety of open and closed spaces and varying degrees of freedom for locomotion. Indoor environments included the "Pinakothek der Moderne" (the Munich modern art museum), a university office building, the main representative lecture building of the University of Munich, and the University Hospital. Outdoor locations included a local forest, Munich's "English Gardens," a residential area ("urban"), small medieval alleys around Munich's "Hofbräuhaus," two large open squares in Munich (Odeonsplatz and Königsplatz), as well as a beach and a

Figure 3. Example movies. Representative 4 s excerpts from each of the 12 environments. *Top row:* Indoor; *middle and bottom row:* Outdoor. In each panel the top represents the gaze camera, the bottom the head camera. Videos have been downsampled and compressed for display purposes, full resolution videos are available from the authors. (Movie should run on most common players and cycle through all 12 environments in 4 s intervals.)

Location	Minutes	Frames
Indoor		
Office	29	44433
University	28	42818
Hospital	65	98509
Pinakothek (museum)	44	66145
Outdoor		
Forest	80	121282
English Gardens	25	38240
Urban	114	172036
Hofbräuhaus	15	23737
Odeonsplatz	54	81369
Königsplatz	25	37555
Desert	72	108369
Beach	9.5	14304

Table 1. Environments and recording times.

desert in Southern California. Recording times within each environment are reported in Table 1. Participants were instructed to behave naturally and freely explore their environment without paying specific attention to the recording equipment. In two environments (in the forest and in front of the Hofbräuhaus), data of one subject each had to be excluded due to recording errors. All procedures conformed with national and institutional guidelines for experiments with human subjects and with the Declaration of Helsinki.

Selection of frames

To restrict analysis to periods of no or slow movements, we estimated gaze and head velocity, as described earlier (Einhäuser et al., 2007). The shift between subsequent frames of each camera is defined by the peak of their 2D cross-correlation in the central region of 256 × 256 pixels. For computational efficiency, images are subsampled at a linear factor of 2, which results in a lower resolution limit for velocity estimation of 0.24°/frame or (5.8°/s) along each axis. Slow or no movement is defined throughout this paper as absolute velocities smaller or equal sqrt(2) pix/frame (8.0°/s).

Since fast head and gaze movements may cause motion blur in the respective camera, only frames with movements slower than 8°/s ("slow frames") are used for analysis. As observers and their gaze move slowly through their environment compared to the 25-Hz sampling rate, subsequent video frames are not independent. To achieve an independent sampling from both cameras, we estimated the length of "slow episodes" that have a successive number of slow head and gaze frames. To not underestimate the length of slow episodes, we applied a median filter of 5 frames to reduce salt-and-pepper noise in the velocity signals. As expected, we find that slow episodes of head are typically longer than slow episodes of gaze. For both cameras, however, the large majority of episodes (80%) are shorter than 20 to 25 frames in most sessions and shorter than 30 frames in all sessions. We therefore selected 1500 slow frames of each movie at random, and whenever two thus selected frames are closer than 2 s (or 50 frames), one of them was randomly chosen and excluded. This leaves about 1000 head and gaze frame pairs in each environment, which are free of motion artifacts and mutually independent.

Analysis of power spectra

We computed the power spectrum of each analyzed frame in the 256 × 256 pixel central region using Matlab's *fft2* function (Mathworks, Natick, MA). To reduce boundary effects, we applied a 256 × 256 circular Hann window to the mean corrected central region before calculation of the power spectra.

Power spectra in natural scenes typically follow a power law (Ruderman & Bialek, 1994). That is, the functional dependence of power P on spatial frequency $f = (f_x, f_y)$ in a particular 1D projection is approximately of the form $P(f) = Af^{\alpha}$. Here we follow Torralba and Oliva's (2003) work on early scene categorization of natural images by differences in the steepness of this power fall. In particular, we consider the horizontal $f = (f_x, 0)$ and the vertical $f = (0, f_y)$ directions. For these 2 directions, an exponent α and a scaling factor A are computed by fitting a linear function in the logarithmic representation $\log P = \alpha \log f + \log A$ for $f > 0$.

Convexity of power spectra

Additionally, we described how uniformly the observed power is distributed among (all) the oblique frequencies by computing the convexity of the region covered by the 5% largest values of the power spectra. Convexity is defined as the proportion of a region's area to the area of its polygonal convex hull (the area of the smallest convex polygon that can contain this region). If the 5% largest values of the power spectra are distributed equally among frequencies of all directions, the region of the power spectra covered by the 5% largest values and its convex hull are both circular and of identical size, hence convexity is 1. In contrast, if power is more selectively distributed to individual orientations such as the cardinals, the top 5% area shows a more concave (or "star-shaped") form which does not fully cover the area of its then diamond shaped convex hull, and convexity is reduced. Hence, convexity is higher if power is distributed more equally among all orientations.

Analysis of local visual features

We calculated maps for various local features at eight different spatial scales of visual angle (0.25°, 0.5°, 1°, 1.5°, 2°, 2.5°, 3°, 4°) on both gaze- and head-aligned movies. For analysis, RGB images were transformed to

the physiologically defined Derrington–Krauskopf–Lennie (DKL) color space (see below) and intensity features were calculated in the luminance dimension of DKL space.

Mean luminance

Mean luminance (ML) in isotropic regions is calculated as a Gaussian low-pass filter, convolving frame I with a Gaussian kernel

$$\mathrm{ML} = I * G, \tag{1}$$

where

$$G = e^{-\left(\frac{x^2}{2\sigma^2}+\frac{y^2}{2\sigma^2}\right)}. \tag{2}$$

Different spatial scales were defined by the full width at half maximum (fwhm) of the Gaussian kernel, such that the standard deviations σ is given by

$$\sigma = \frac{2\sqrt{2\ln(x)}}{\mathrm{fwhm}}. \tag{3}$$

Luminance contrast

In line with earlier eye-tracking studies (e.g., Einhäuser & König, 2003; Reinagel & Zador, 1999) and as the straightforward generalization of two-point contrast, we defined luminance contrast of a local region as the standard deviation of luminance in an isotropic patch around a point, normalized by the mean luminance of the image:

$$\mathrm{LC} = ((I^2 * G) - (I * G))/\langle I \rangle. \tag{4}$$

LC was calculated at the same spatial scales than mean luminance by adjusting the size of the isotropic patch with the fwhm of the Gaussian kernel.

Texture contrast

We define "texture contrast" as a higher-order variation of luminance contrast without any further model assumptions as a canonical generalization of our definition of luminance contrast. Texture contrast in a region is the standard deviation of luminance contrast in an isotropic region around a point:

$$\mathrm{TC} = (\mathrm{LC}^2 * G) - (\mathrm{LC} * G). \tag{5}$$

We calculated TC at spatial scales one and two times larger than the underlying LC scale, named TC and TC2, respectively.

Color contrasts

Analysis of color contrasts requires an independent coding of hue and intensity, such as in the well-known "hue-saturation-value" (HSV) color space. To ensure consistency with physiological and psychophysical studies, here we compute color contrasts in a physiological color space, which is based on the relative excitations of the three cone types (L, M, and S) in the primate retina, the Derrington–Krauskopf–Lennie (DKL) color space (Derrington, Krauskopf, & Lennie, 1984). It is spanned by two color axis "constant blue" or "yellow-blue (YB)" (defined by the difference between L and M cone excitations) and "tritanopic confusion" or "red-green (RG)" (defined as L + M − S cone excitations), and a "luminance" axis. We examined the contribution of the two color-opponent processes to overt attention independently by measuring color contrast separately for both cardinal color axes. As in the definition of luminance contrast, we defined the RGC and YBC color contrasts in an isotropic local region as the standard deviation of the chromatic content of the region. Different spatial scales again were defined by adjusting the FWHM of the isotropic kernel.

Bar-ness

We described (edge) isotropy of local regions in a measure of "bar-ness," using standard structure tensor methods (Jähne, 1997). The structure tensor matrix is defined as

$$S = \begin{bmatrix} J_{xx} & J_{xy} \\ J_{xy} & J_{yy} \end{bmatrix}, \tag{6}$$

with

$$J_{xx} = I_x^2 * G \tag{7}$$

$$J_{yy} = I_y^2 * G \tag{8}$$

$$J_{xy} = I_x I_y * G, \tag{9}$$

where gradient images I_x and I_y are obtained by Sobel filtering of individual frames. The spatial scale is defined via the fwhm of the isotropic Gaussian G. Eigenvectors of S represent the direction of longest and smallest axis of inertia, and their eigenvalues λ_1 and λ_2 describe the extent

of the gradient in the respective dimension. The coherence of local orientation, i.e., the ratio of the difference to the sum of the extents of the gradients on the smallest and longest axes of inertia, is given by:

$$\text{Bar-ness} = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} = \frac{(J_{yy} - J_{xx})^2 + 4J_{xy}^2}{(J_{xx} + J_{yy})^2}. \quad (10)$$

To avoid numerical problems due to near 0 values of the denominator, the measure was regularized by adding Matlab's *eps* (2^{-52}) to the denominator. The measure characterizes how well a patch represents an oriented structure (or a bar) and is referred to as "bar-ness." Bar-ness is 0 for an isotropic structure and 1 for a perfectly oriented structure.

Color bar-ness

To describe the extent of local orientation in the color channels, we applied the definition of bar-ness also to both color channels at the same spatial scales.

Since the current study evaluates the spatial distributions of local features, absolute feature values were mean normalized by *z*-transformation for individual frames. Hence, feature values below report the intensity of a feature relative to its spatial neighbors.

Spatial distribution of features in gaze- and head-centered coordinates

To compare the spatial distribution of local features in gaze- and head-centered coordinates at a scale large compared to the features themselves, we computed feature maps for each feature in each of the 1000 selected slow frames per camera. Averaging these maps for each environment and camera reflects the topography of this feature in the respective coordinate system. We quantified the average feature maps by the position of the peak value. In addition we computed the anisotropy of the feature distribution.

Peak detection

We used morphological image processing to detect the position of the feature peak. Peak position is computed by extended-maxima transformation (using Matlab's *imextendedmax* function) as the position of the regional maxima out of 8-connected pixel neighborhoods containing only values in the top 2% and with external boundary pixels that have lower values. Before peak detection, mean feature maps were median filtered in an 8×8 pixel neighborhood for removal of potential dust and scratch noise. For one observer, in one of our outdoor and indoor sessions, a grain of sand contaminates a peripheral region in the head camera. For analysis this region, which covers less than two percent of the image area, was masked and excluded on both cameras in the respective sessions. Peak detection was not affected by this procedure in any feature or environment.

Anisotropy

The anisotropy of a feature topography was computed in the ellipse that has the same second-order moment than the region covered by 60th to 90th percentile of the feature map (using Matlab's *regionprops* function). Anisotropy is calculated as the eccentricity of the ellipse, i.e., the ratio of the distance between the foci of an elliptical region and the length of its major axis. It is 0 for a fully circular topographical region and 1 for an elliptical region that resembles a line segment and quantifies the isotropy of the feature topography in gaze- and head-centered coordinates.

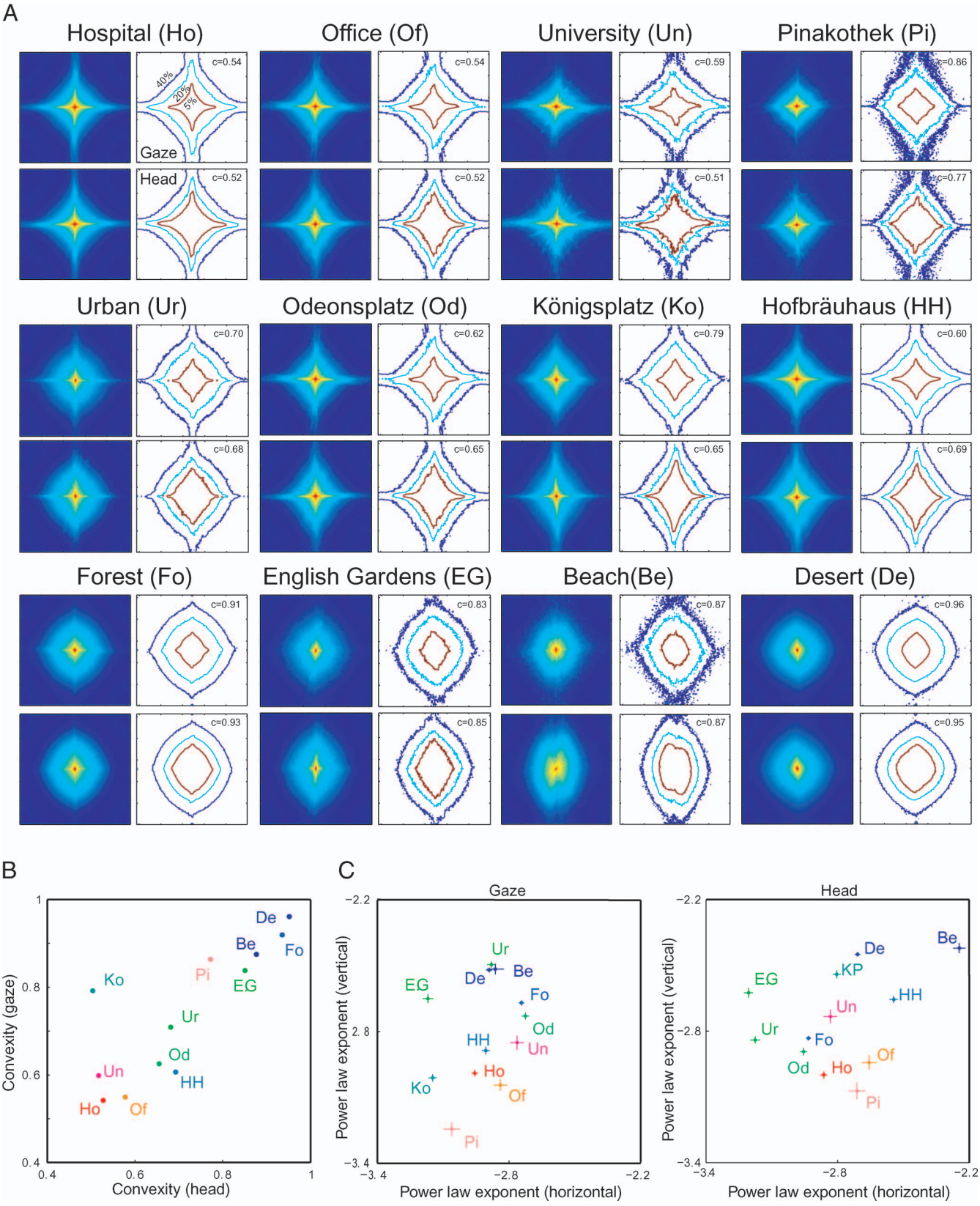
Results

Environments

We simultaneously recorded videos of gaze- and head-centered visual stimuli attended by a human viewer during natural exploration. The head-centered recordings capture the statistical properties of the environment upon which gaze guidance may operate. As a first step, we characterized the 12 different environments on the basis of their average power spectra (Figure 4A). By visual inspection, in the indoor environments (Figure 4A, top row), which are spatially constrained by walls and ceilings, a majority of power is confined to the cardinal axis, leading to a

Figure 4. (A) Average power spectra for all environments. The four subpanels for each environment show the average power spectrum over a sample of about 1000 slow frames for the gaze camera (top left) and head camera (bottom left) in logarithmic scale, from red to blue. DC component in the center, frequencies linearly spaced from DC to Nyquist limit of 5.5 cycles/deg, axis rotated such that $P(f_x)$ runs along the horizontal axis. Plots on the right show percentile contour lines (5%, 20%, 40%, excluding the DC component) for the respective left hand side subpanel power spectra. Environments sorted as in Figure 2. (B) Convexity of top 5% area of power spectra (red line in Figure 4A). Convexity is 1 if power is distributed equally among (all) the oblique frequencies and the top 5% is circular; it is reduced if power is distributed more selectively to orientations close to the cardinal orientations and the top 5% area is more concave (or "star-shaped"). (C) Fitted exponents α for power spectra (horizontal and vertical), mean and standard error for each environment (color coded). *Left:* Gaze camera; *right:* Head camera.

pronounced anisotropy in their power spectra. This is true with the exception of the recordings in the modern art museum "Pinakothek." In contrast, outdoor environments seem to fall into two classes. On the one hand, anisotropy is weakest for "open" environments ("Beach," "Desert," "Forest," and "English Gardens"). On the other hand, outdoor environments containing streets and open city squares with large buildings ("Urban," "Hofbräuhaus,"



"Odeonsplatz," and "Königsplatz") behave more like indoor environments. For the remainder of this analysis, we treat these outdoor categories separately and refer to them as "open" and "closed," respectively. We quantified the isotropy of the average power spectra by measuring the convexity of the central area containing the 5% largest values (inner line in Figure 4A). Confirming the visual impression, we find that convexity is smallest (i.e., anisotropy is largest) for indoor environments (0.54 ± 0.03, mean ± *SD*; Figure 4B, with the exception of the Pinakothek). Convexity is largest for the open outdoor environments (0.90 ± 0.04, mean ± *SD*), which is significantly different from the indoor mean ($p = 0.0009$, $t(5) = 11.21$). In contrast, the convexity of closed outdoor environments is closer to that of indoor environments (0.63 ± 0.08), and the two are not significantly different ($t(5) = 1.70$, $p = 0.14$). The isotropy of the spectral signature in the modern art museum (0.77) is more similar to the open outdoor environments than to the indoor environments. As the recordings in the Pinakothek contain a large number of close-up sequences of paintings presented on a wall, this is in line with earlier findings reporting more isotropic spectral signatures with closer scene background (Torralba & Oliva, 2003).

The differences between environment classes are also reflected in the average exponent of the power law fits for the head camera (Figure 4C, right): the indoor environments tend to have larger (more negative) exponents in the vertical direction (horizontal: mean −2.76 ± 0.09, vertical: −2.89 ± 0.14) than the open outdoor environments (horizontal: mean −2.77 ± 0.40; vertical: −2.58 ± 0.18, *t*-test $t(6) = -2.97$, $p = 0.02$). However, the vertical exponents of "closed" outdoor environments (horizontal mean −2.86 ± 0.26, vertical mean −2.73 ± 0.16) fall between these but are not distinguishable from the indoor category (vertical: $t(6) = -1.86$, $p = 0.11$). Interestingly, horizontal exponents show no difference between the groups (ANOVA, $F(2, 9) = 0.12$, $p = 0.88$). These results show that our recorded environments do categorically differ from each other and that the extremes of these categories (indoor versus open outdoor) are separable on the basis of their power spectra. The characterization into "indoor" and "outdoor" alone, however, does not present a clear dividing line with respect to the power spectra of real-world scenes.

Qualitative differences between the form of average power spectra in gaze and head cameras are minute (Figures 4A and 4B). Although the difference between fitted exponents for gaze and head are sometimes of the same order as the differences between environments, the general pattern is preserved: open outdoor environments have small vertical exponents, indoor environments have larger ones, and the other outdoor environments are in between (Figure 4C, middle). This shows that the global characteristics of the individual frames are not influenced by changes in gaze. In other words, the analyzed field of view is sufficiently large that eye and head frames are identical with respect to their global spatial frequency statistics. Consequently, comparing the spatial distribution of local features between eye and head cameras will allow us to investigate their effect on the adjustment of gaze and attention.

Topography of local features at head and gaze

We adopt the common definition that a feature is "salient" if it is elevated at the center of gaze. There can be two explanations for an increase in the saliency of a feature: either the feature is already centered in head coordinates (i.e., the stimuli upon which eye-movements operate have a central bias), or eye-movements actively contribute to the elevation of the feature at the center of gaze. In the former case, feature maps in head- and gaze-centered coordinates (i.e., cameras) should be similar; in the latter case, the peak of the gaze-centered map is predicted to be more centrally located.

Recordings in the head camera show that local image features are typically not uniformly distributed, but biased in spatial location. In the example environment "office" (Figure 5, bottom panels), head-aligned maps show pronounced peaks for all features, but the three bar-ness features. The peaks for these 6 features (ML, LC, TC1, TC2, RGC, and YBC) always fall above the horizontal midline and have an average distance to the center of 8.6° ± 1.5° (observer 1, mean ± *SD*) and 8.2 ± 1.73 (observer 2). Unlike in most experiments with static stimuli, eye-in-head movements therefore operate on feature maps that have an upward bias rather than a central one. If central biases in fixation were just due to resetting eyes in orbit, as mentioned as one possible alternative in Tatler (2007), maps in gaze-centered coordinates should exhibit the peak at the same position. Instead, the peaks are closer to the center in the gaze-centered camera (Figure 5, top panels), averaging (over the same 6 features as above) to 3.9° ± 1.6° in observer 1 and 5.0° ± 2.6° in observer 2. As this requires the eyes to look upward relative to the head, eye-in-head movements actively center these stimulus features.

Can this effect be explained by a mere misalignment of the head camera? In addition to the robustness of the setup, which already makes a 9° offset very unlikely, misalignment would only shift the peaks but would not change their shape. Here however we observe that peaks are not only more central in gaze, but also more narrow and isotropic: contour lines in head-centered feature maps are elongated along the horizontal axis (anisotropy over the six features observer 1: 0.85 ± 0.07; observer 2: 0.93 ± 0.02), while features in gaze-centered coordinates are more equally distributed along the spatial directions (observer 1: 0.53 ± 0.16; observer 2: 0.59 ± 0.29). This also holds for the more uniformly distributed bar-ness features, which show a pronounced peak only for gaze-centered maps. This increased isotropy for gaze-centered features again argues in favor of an active selection process through eye-in-head movements. In the case of the office environment example, these data demonstrate the active role of eye-in-head movements in centering salient features.

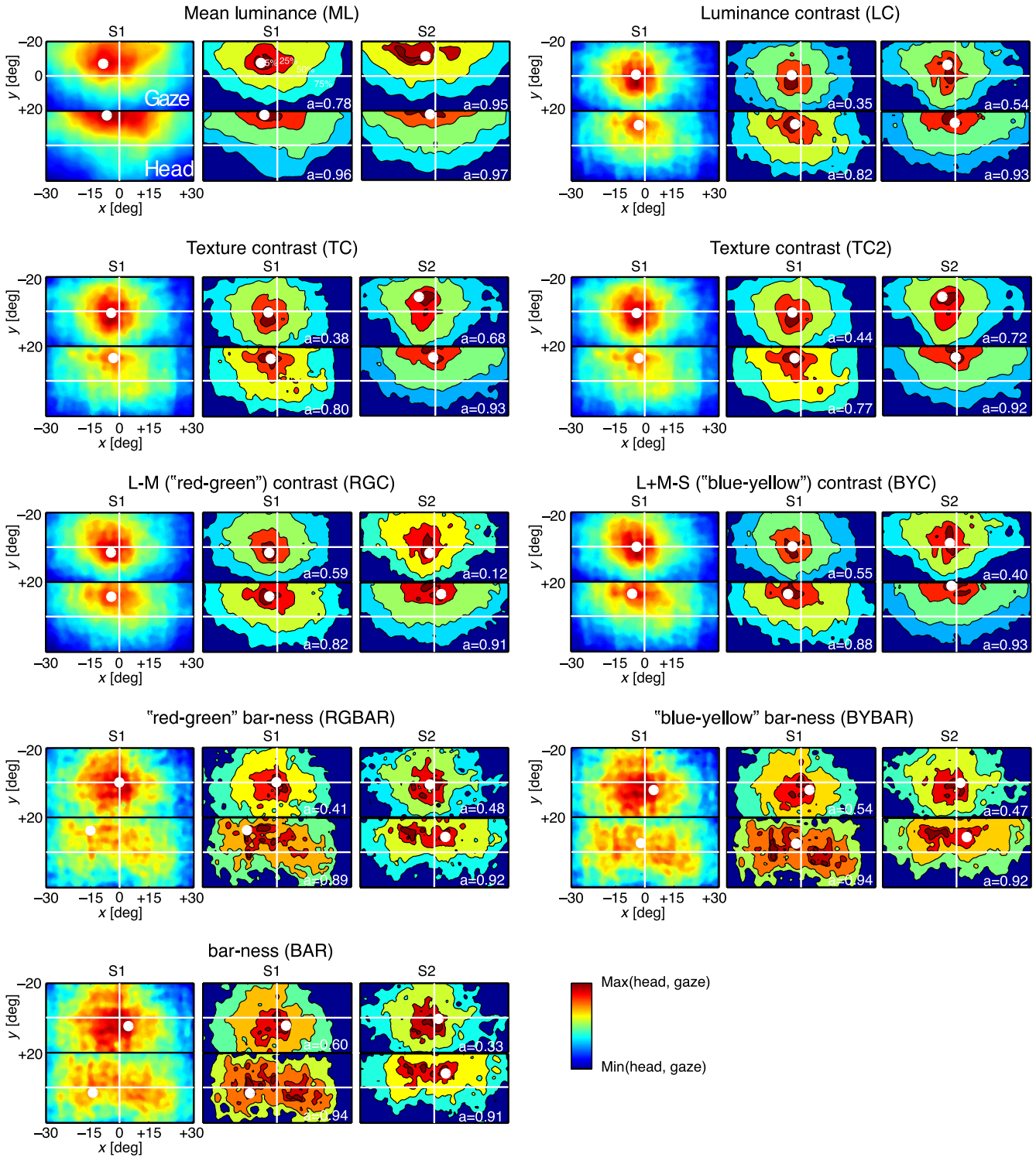
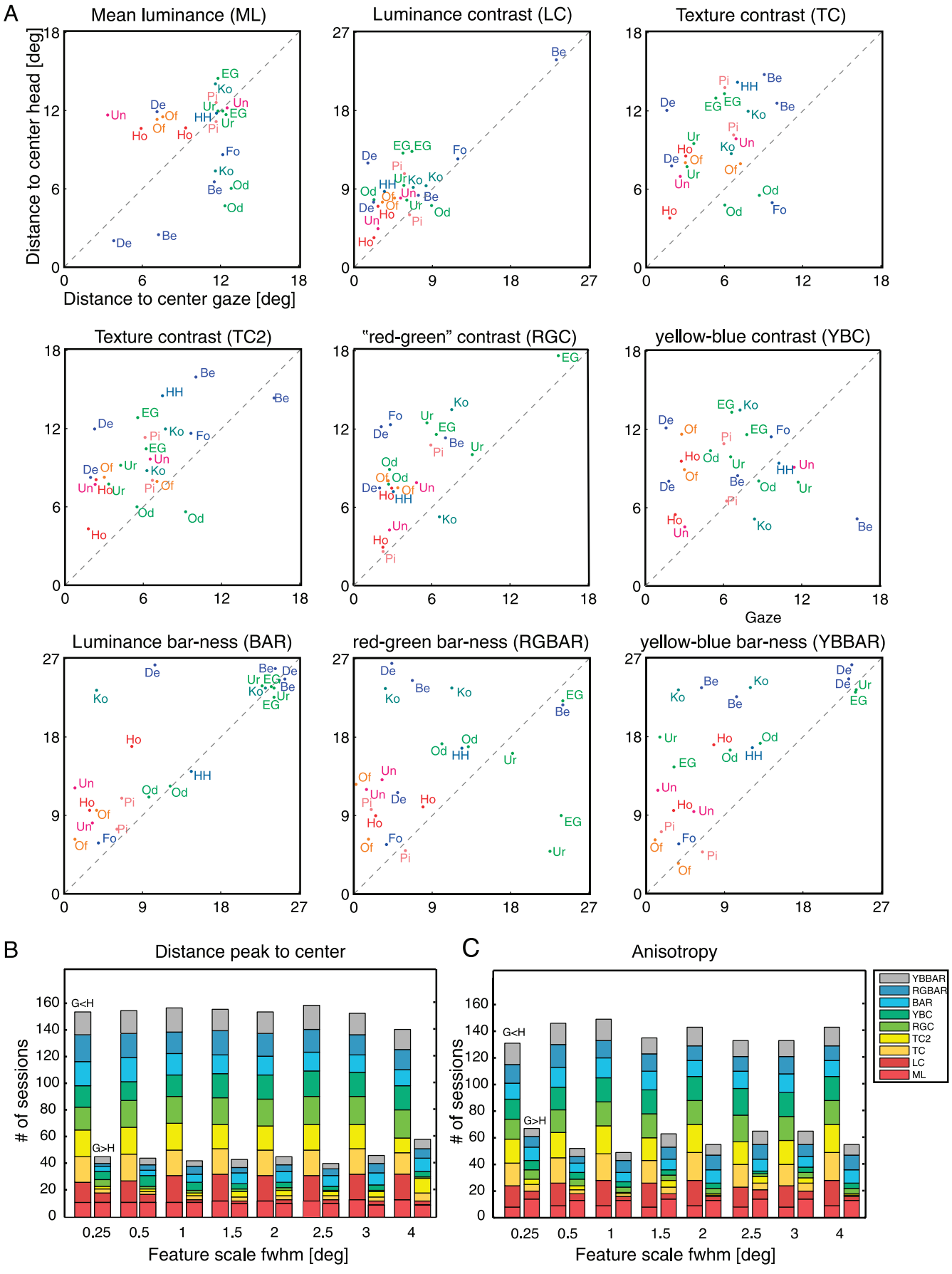


Figure 5. Topography of feature distributions. Feature maps of all analyzed local features for office environment at a feature size of 1° fwhm. Elevation at a spatial coordinate depicts the mean intensity of the feature over a sample of about 1000 slow frames. In each subpanel: *top*: Gaze camera; *bottom:* Head camera; *left:* Raw feature maps for one observer; *middle:* Corresponding percentile map (5%, 25%, 50%, 75%); *right*: Percentile map for the other observer. White dots represent center of image; white diamond, peak of the map. Anisotropy of the ellipse best fitting the region covered by values between the 60th and 90th percentile is given in the measure a. Features from left to right and top to bottom: Mean luminance (ML), luminance contrast (LC), texture contrast with first and second order filters on same scale (TC), texture contrast with second order filter double the first order scale (TC2), red-green (L − M) contrast (RGC), blue-yellow (L + M − S) contrast (BYC), red green bar-ness (RGBAR), yellow blue bar-ness (BYBAR), luminance bar-ness (BAR). Bar-ness describes the extent of oriented bars within a local patch irrespective of their directions.



Do these observations generalize to other environments? Analyzing all features at a scale of 1° fwhm, we note that the distance between the peak of a head-centered feature map and its center varies substantially between environments (vertical axes in Figure 6A). The spatial bias in head-centered coordinates thus depends on the environment. Interestingly, the data of the two subjects within the same environment is typically quite similar compared to this inter-environment variability. For each feature, we quantified this observation by comparing the peak-to-center distance within the same environment (inter-observer difference) and between different environments (inter-environment difference). We find the mean inter-observer difference to be smaller than the mean inter-environment difference in all of the 9 features tested (inter-observer/inter-environment—ML: 2.2°/3.0°; LC: 3.0°/3.7°; TC1: 1.6°/2.7°; TC2: 2.2°/3.3°; RGC: 3.3°/4.0°; YBC: 2.9°/3.8°; BAR: 4.8°/9.8°; RBAR: 4.5°/8.5°; YBAR: 7.5°/8.1°). This again provides evidence against misalignment artifacts since there is no reason why any putative misalignment of the head camera with the head should depend more on the environment than on the observer. More importantly, it stresses the influence of the environment on head-centered statistics as compared to idiosyncratic factors.

Despite the large variability in the position of feature peaks in the head camera, the peak in the gaze camera is closer to the center in most environments (Figure 6A). For each feature, we assess whether the fraction of the 22 recording sessions (10 environments × 2 subjects + 2 environments × 1 subject) in which the gaze peak is closer to the center than the head peak, is significant. To obtain a robust measure that only considers the direction of the effect (whether most points falls above or below the diagonal in Figure 6A), but not its size, we used the sign test. We find that the fraction of sessions for which gaze peaks are more central is significant for all features except mean luminance (ML: $p = 0.97$; BAR, YBBAR, RGBAR: $p = 0.05$; LC, TC2, RGC: $p = 0.0001$; YBAR: $p = 0.004$; TC1: $p = 0.0008$). This pattern holds not only for the 1° fwhm features reported up to now but also across the full range of spatial scales tested (Figure 6B). This demonstrates that eye-in-head movements robustly center the gaze on salient features.

Does the second hallmark of active selection of salient features by eye-in-head movements—the increased isotropy for gaze-centered as compared to head-centered feature maps—also generalize from "office" to all environments? We quantify anisotropy by measuring the eccentricity of an ellipse fitted to a fixed percentile range (60% to 90%). For each of the 1° fwhm features except ML and the three bar-ness features (BAR, RBAR, YBAR), we find a significant fraction of environments in which the feature map's peak is more isotropic in gaze than in head-centered coordinates (ML, RGBAR: $p = 0.523$; BAR: $p = 0.286$, LC: $p = 0.016$; TC, RGC, YBC: $p = 0.004$, TC2: $p = 0.0008$). Again, this result generalizes over all scales tested (Figure 6C). In summary, feature maps have more centralized and isotropic peaks in gaze-centered as compared to head-centered coordinates. Under real-world recording conditions, salient features are therefore not a mere consequence of environmental or heading biases, but of an active selection process by means of eye-in-head movements.

Figure 6. Gaze centers features. (A) Euclidian distance of peak in feature map to center for head (*y*-axis) and gaze (*x*-axis) in all features (1° as in Figure 4). Environment color coded, two observers for most environments. Note that in all plots most points fall above the diagonal, i.e., the feature peaks more central in gaze (G < H) than in head-centered coordinates (G > H). (B) Summary of results for different feature scales (fwhm). *Left bar*: Feature more central in gaze (G < H); *right bar*: Feature more central in head (G > H), counts summarized over different features (color coded) and environments (24 per feature). (C) Analogous plot for anisotropy of the ellipse best fitting the region covered by values between the 60th and 90th percentile. *Left bar*: Feature more isotropic in gaze (G < H); *right bar*: Feature more isotropic in head (G > H). Note that in most environments features are distributed more isotropic in the gaze than in the head-centered coordinates.

Discussion

In the present study, we recorded a large amount of video data to compare the spatial distribution of stimulus features in head- and gaze-centered coordinates during free natural exploration behavior. When the environments were characterized by their power spectra, two distinct classes of outdoor scenes emerged: those spatially constrained by large buildings which show anisotropic spectral signatures with emphasized horizontal and vertical spatial frequencies similar to indoor environments. This is in contrast to open environments (beach, desert, etc.) which distribute spectral power more equally along frequencies of all orientations. The modern art museum Pinakothek, which contains a large fraction of close-up views on drawings, also showed more isotropic power spectra, demonstrating a dependence of the spectral characteristics on the distance from the scene (Torralba & Oliva, 2003). We found little difference in global power spectra between gaze- and head-centered recordings. However, local salient image structures were shown to be actively selected by eye-movements, relative to the environment-dependent feature biases in a head-centered coordinate frame. Hence, we demonstrate that salient features, i.e., features elevated at the center of gaze, are not only a mere correlative consequence of the stimulus and fixation biases induced by typical laboratory setups, but can also be found in a natural setting.

It is important to note that the present operational definition of saliency, elevation of feature values at the center of gaze, does not imply that those features indeed drive gaze or attention causally, as has been pointed out earlier (Carmi & Itti, 2006; Einhäuser & König, 2003). However, we extend previous studies that addressed stimulus statistics at the center of gaze (Krieger et al., 2000; Mannan et al., 1996, 1997; Privitera & Stark, 2000; Reinagel & Zador, 1999) in three respects. First, those studies are restricted to saccadic eye-movements, i.e., assume that gaze allocation constitutes a sequence of static fixations interrupted by large volitional saccades. However, under real-world conditions, compensatory eye-movements play an important role and the input to the human retina cannot be adequately modeled using fixations and saccades alone (Einhäuser et al., 2007). Second, most laboratory setups necessitate that the field of view is restricted and the head is fixed, thereby potentially introducing effects of screen boundaries, which may center the gaze relative to the screen. Our finding that gaze centers feature peaks that are off-center in head coordinates argues against a tendency to re-center eyes in their orbit during natural behavior, similar to findings of Vitu, Kapoula, Lancelin, and Lavigne (2004) during reading. Our data thus confirm the suspicion of Tatler (2007) that the stimulus-independent central bias of fixation observed in his study is best explained as an artifact of centering the eyes relative to the display screen. Consequently, our data provide further support for an important implication pointed out by Tatler: the central fixation bias requires careful compensation in monitor-based studies, which highlights the prospects of unrestrained eye-tracking experiments with full-field-of-view. Third, the aforementioned studies involve the presentation of pre-selected image material. If this includes photographs taken by a human, they typically already exhibit a central bias in features of the order of the expected effect (Tatler et al., 2005). If the images are obtained in a less anthropocentric manner, e.g., from a car (van Hateren & Ruderman, 1998) or in head coordinates of animals such as cats (Betsch, Einhäuser, Körding, & König, 2004) or owls (Ohayon, Harmening, Wagner, & Rivlin, 2008), different biases with respect to environment or perspective may be introduced. In fact, we here demonstrate that for a realistic assessment of the role of eye-movements relative to head-centered coordinates, stimuli *should* be biased. Most features in the head camera are *not* uniformly distributed. Instead the distribution is determined by the environment, but also has some commonalities, such as the peak above the midline and the anisotropy of the distribution. Only relative to this realistic, environment-dependent baseline can the role of eye-movements in gaze allocation be assessed.

In sensory development, many response properties of cortical cells have been argued to optimize spatial, temporal, or spatiotemporal objective functions of natural input. In vision, such models by now cover the entire ventral stream: optimizing sparseness and/or temporal coherence under natural scenes or videos yields properties of V1 simple cells (Bell & Sejnowski, 1997; Olshausen & Field, 1996; van Hateren & Ruderman, 1998), V1 complex cells (Berkes & Wiskott, 2005; Einhäuser, Kayser, König, & Körding, 2002; Körding, Kayser, Einhäuser, & König, 2004), IT-like invariant object representations (Einhäuser, Hipp, Eggert, Körner, & König, 2005; Stringer & Rolls, 2002), and—beyond vision—even hippocampal place fields (Franzius, Sprekeler, & Wiskott, 2007; Wyss, König, & Verschure, 2006). Similarly, several physiological and psychophysical effects (e.g., the distance-size illusion; Howe & Purves, 2002) and the development of spatial representations in the visual system (Baddeley, 1997) have been explained by adaptation to stimulus statistics. Even though some of these studies employ head-centered recordings (of cats or robots) during free behavior, they typically do not address the role of eye movements (but see Li & Clark, 2004). In addition to the obvious influence of eye movements on temporal statistics, we here also demonstrate that recordings of natural stimuli, even if obtained by a head-fixed camera, do not faithfully represent the spatial statistics of human visual input either. As has been noted in a pointed remark of Pinto, Cox, and DiCarlo (2008), using stimuli that do not properly reflect essential properties of natural input can be "seriously misleading, potentially guiding progress in the wrong direction." Our findings show that for a truthful recording of natural human input, head-fixed recordings are not sufficient, and gaze-centered stimuli should be recorded in a situation where eyes, body, and head can freely move.

Here we did not use an explicit task, but asked observers to behave naturally. Surprisingly, the feature maps of the two observers in a given environment are nevertheless remarkably similar. In fact, even the distance between the maps' peaks to the stimulus center in head and gaze alone already would allow a coarse categorization of the environments (Figure 5A). Given the classical observations on the importance of task when looking at pictures (Buswell, 1935; Yarbus, 1967), the large body of literature studying gaze allocation for specialized tasks (Furneaux & Land, 1999; Land & Hayhoe, 2001; Land et al., 1999; Pelz et al., 2000) and the recent results on the importance of task while navigating virtual reality (Rothkopf et al., 2007), it is interesting to speculate that a particular environment may introduce similar implicit tasks in different observers even under free exploration. Because of the heterogeneity of a single environment with respect to local features (reflected in the non-uniformity of the head maps), one might speculate that these are rather generic tasks or rather high-level objectives, such as finding an open path, walking safely on uneven terrain, recognizing landmarks, or avoiding collisions with objects. Following this speculation raises the question of what it is that purely bottom-up studies of scene perception—in the laboratory or during free exploration—actually measure. Even if gaze were entirely task-driven, salient features are expected at gaze if non-

salient regions contain little task-relevant information worth inspecting. Indeed, in natural scenes, interesting regions and objects are correlated with saliency (Elazary & Itti, 2008). Hence, on the one hand, the idea of purely bottom-up control of scene perception may be overly simplistic, and implicit higher-level tasks may always dominate gaze control. On the other hand, it is conceivable that bottom-up models describe a "common denominator" over all possible generic tasks a given environment may invoke. Conditioned on the task, i.e., if the task is known or made explicit, pure bottom-up models must "fail" (e.g., Henderson et al., 2006), as they provide little or no additional information. If, however, all tasks are unknown or implicit, the same bottom-up models still may have predictive power. Incorporating the task (top-down information) into more and more sophisticated bottom-up models, as has already been done for saliency maps (Navalpakkam & Itti, 2007), is thus a promising starting point to bridge the gap between unconditional (i.e., "bottom-up") and task-conditioned (i.e., "top-down") models of attention and perception. Here, as a first step in a truly natural setting, we quantified stimulus statistics at the center of gaze without an explicit task. Future research addressing the effect of specific, explicit tasks in our natural setting will then provide further insight into the interplay between goal-directed behavior, environmental constraints, and stimulus properties in the allocation of gaze.

Acknowledgments

This work was financially supported by the Bavarian-Californian Technology Center (BaCaTeC), Bayerische Forschungsstiftung (DOK-88-07 to JV), in part within the DFG excellence initiative research cluster Cognition for Technical Systems-CoTeSys, and the EU project "Perception on Purpose" (POP). We would like to thank Thomas Dera for his work on the eye-tracker algorithms. We thank Cliodhna Quickley, Benjamin Tatler, and an anonymous reviewer for their comments on earlier versions of the manuscript.

Commercial relationships: none.
Corresponding author: Frank Schumann.
Email: fschuman@uni-osnabrueck.de.
Address: Institute of Cognitive Science, University of Osnabrück, Albrechtstrasse 28, 49069 Osnabrück, Germany.

References

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science, 21,* 351–372. [Article]

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46,* 2824–2833. [PubMed]

Ballard, D. H., Hayhoe, M. M., Li, F., & Whitehead, S. D. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 337,* 331–338. [PubMed]

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 37,* 3327–3338. [PubMed]

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision, 5*(6):9, 579–602, http://journalofvision.org/5/6/9/, doi:10.1167/5.6.9. [PubMed] [Article]

Betsch, B. Y., Einhäuser, W., Körding, K. P., & König, P. (2004). The world from a cat's perspective–statistics of natural videos. *Biological Cybernetics, 90,* 41–50. [PubMed]

Brandt, T., Glasauer, S., & Schneider, E. (2006). A third eye for the surgeon. *Journal of Neurology, Neurosurgery, and Psychiatry, 77,* 278. [PubMed]

Buswell, G. T. (1935). *How people look at pictures. A study of the psychology of perception in art*. Chicago, IL: The University of Chicago Press.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46,* 4333–4345. [PubMed]

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (vol. 20), Cambridge, MA: MIT Press.

Chajka, K., Hayhoe, M., Sullivan, B., Pelz, J., Mennie, N., & Droll, J. (2006). Predictive eye movements in squash [Abstract]. *Journal of Vision, 6*(6):481, 481a, http://journalofvision.org/6/6/481/, doi:10.1167/6.6.481.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology, 357,* 241–265. [PubMed] [Article]

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics, 93,* 79–90. [PubMed]

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience, 15,* 475–486. [PubMed]

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual

attention? *European Journal of Neuroscience, 17,* 1089–1097. [PubMed]

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1–19, http://journalofvision.org/8/2/2/, doi:10.1167/8.2.2. [PubMed] [Article]

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., et al. (2007). Human eye-head co-ordination in natural exploration. *Network, 18,* 267–297. [PubMed]

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3):3, 1–15, http://journalofvision.org/8/3/3/, doi:10.1167/8.3.3. [Article]

Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology, 3,* e166. [PubMed] [Article]

Furneaux, S., & Land, M. F. (1999). The effects of skill on the eye-hand span during musical sight-reading. *Proceedings of the Royal Society B: Biological Sciences, 266,* 2435–2440. [PubMed] [Article]

Glasauer, S., Schneider, E., Jahn, K., Strupp, M., & Brandt, T. (2005). How the eyes move the body. *Neurology, 65,* 1291–1293. [PubMed]

Hamker, F. H. (2006) Modeling feature-based attention as an active top-down inference process. *Biosystems, 86,* 91–99. [PubMed]

Harris, J. M., & Rogers, B. J. (1999). Going against the flow. *Trends in Cognitive Sciences, 3,* 449–450. [PubMed]

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188–194. [PubMed]

Hayhoe, M., Mannie, N., Sullivan, B., & Gorgos, K. (2005, Fall). The role of internal models and prediction in catching balls. In *Proceedings of AAAI 2005 Fall Symposium.*

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2006). Visual saliency does not account for eye-movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain*. Oxford: Elsevier.

Howe, C. Q., & Purves, D. (2002). Range image statistics can explain the anomalous perception of length. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 13184–13188. [PubMed] [Article]

Jähne, B. (1997). *Digital image processing: Concepts, algorithms, and scientific applications*. Berlin: Springer.

Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research, 46,* 2535–2545. [PubMed]

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219–227. [PubMed]

Körding, K. P., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology, 91,* 206–212. [PubMed] [Article]

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13,* 201–214. [PubMed]

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research, 25,* 296–324. [PubMed]

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41,* 3559–3565. [PubMed]

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 69,* 742–744. [PubMed]

Land, M. F., & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience, 3,* 1340–1345. [PubMed]

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28,* 1311–1328. [PubMed]

Lappe, M., Bremmer, F., & van den Berg, A. V. (1999a). Reply to Harris and Rogers. *Trends in Cognitive Sciences, 3,* 450. [PubMed]

Lappe, M., Bremmer, F., & van den Berg, A. V. (1999b). Perception of self-motion from visual flow. *Trends in Cognitive Sciences, 3,* 329–336. [PubMed]

Li, M., & Clark, J. J. (2004). A temporal stability approach to position and attention-shift-invariant recognition. *Neural Computation, 16,* 2293–2321. [PubMed]

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10,* 165–188. [PubMed]

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11,* 157–178. [PubMed]

Mial, R. C., & Tchalenko, J. (2001). A painter's eye movements: A study of eye and hand movement during portrait drawing. *Leonardo, 34,* 35–40.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53,* 605–617. [PubMed] [Article]

Ohayon, S., Harmening, W., Wagner, H., & Rivlin, E. (2008). Through a barn owl's eyes: Interactions between scene content and visual attention. *Biological Cybernetics, 98,* 115–132. [PubMed]

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609. [PubMed]

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123. [PubMed]

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19,* 783–789. [PubMed]

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41,* 3587–3596. [PubMed]

Pelz, J. B., Canosa, R., Babcock, J., Kucharczyk, D., Silver, A., & Konno, D. (2000). Portable eyetracking: A study of natural eye movements. In *Proceedings of the SPIE: Vol. 3959. Human vision and electronic imaging* (pp. 566–583). SPIE.

Perrone, J. A., & Stone, L. S. (1994). A model of self-motion estimation within primate extrastriate visual cortex. *Vision Research, 34,* 2917–2938. [PubMed]

Peters, R. J., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception, 5,* 8.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45,* 2397–2416. [PubMed]

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology, 4,* e27. [PubMed] [Article]

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46,* 1886–1900. [PubMed]

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest. Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 970–982.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10,* 341–350. [PubMed]

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14):16, 1–20, http://journalofvision.org/7/14/16/, doi:10.1167/7.14.16. [PubMed] [Article]

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters, 73,* 814–817. [PubMed]

Rutishauser, U., & Koch, C. (2007). Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision, 7*(6):5, 1–20, http://journalofvision.org/7/6/5/, doi:10.1167/7.6.5. [PubMed] [Article]

Schneider, E., Bartl, K., Bardins, S., Dera, T., Boning, G., & Brandt, T. (2005). Eye movement driven head-mounted camera: It looks where the eyes look. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 3,* 2437–2442.

Schneider, E., Bartl, K., Dera, T., Böning, G., Wagner, P., & Brandt, T. (2006). Documentation and teaching of surgery with an eye movement driven head-mounted camera: See what the surgeon sees and does. *Studies in Health Technology and Informatics, 119,* 486–490. [PubMed]

Stringer, S. M., & Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation, 14,* 2585–2596. [PubMed]

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://journalofvision.org/7/14/4/, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659. [PubMed]

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14,* 391–412. [PubMed]

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786. [PubMed]

van Hateren, J. H., & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences, 265,* 2315–2320. [PubMed] [Article]

Vitu, F., Kapoula, Z., Lancelin, D., & Lavigne, F. (2004). Eye movements in reading isolated words: Evidence for strong biases towards the center of the screen. *Vision Research, 44,* 321–338. [PubMed]

Vockeroth, J., Bardins, S., Bartl, K., Dera, T., & Schneider, E. (2007). The combination of a mobile gaze-driven and a head-mounted camera in a hybrid perspective setup. *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics,* 2576–2581.

Wilkie, R., & Wann, J. (2003). Controlling steering and judging heading: Retinal flow, visual direction, and extraretinal information. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 363–378. [PubMed]

Wyss, R., König, P., & Verschure, P. F. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology, 4,* e120. [PubMed] [Article]

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.